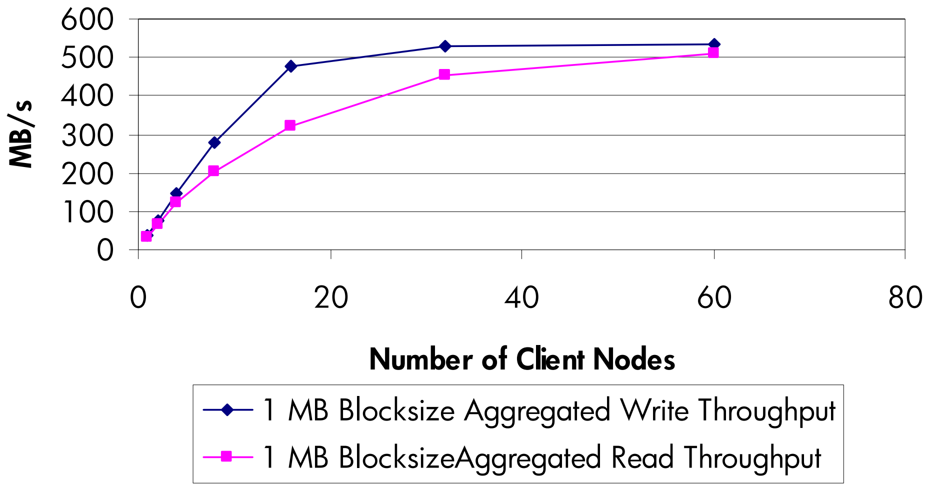
File System Experise @ SCS

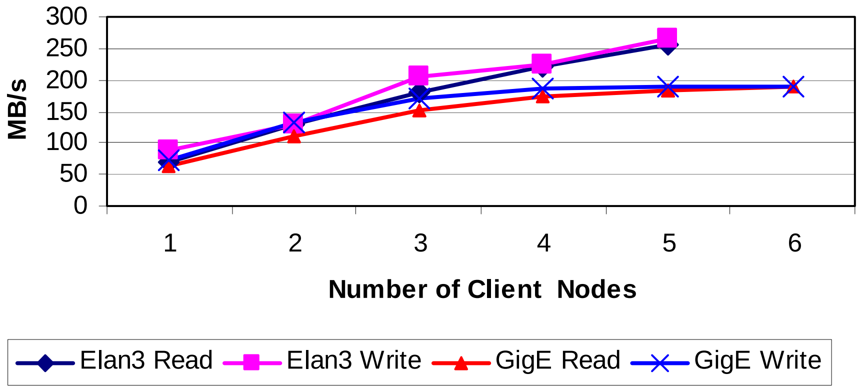
Supercomputing Systems

- Work done since 1999:
  - Re-engineered Petal/Frangipani research code
  - Port of Petal/Frangipani to 2.4 Linux Kernel
    - single client throughput up from 35 MB/s to 102 MB/s (153 MB/s with direct I/O), thanks to our page cache & aggregation logic
  - Design done for fine grain locking, deferred m_time updates, efficient MPI-I/O support, others
  - Distributed test suite, a lot of fixes done

Datarate for streaming I/O on Petal/Frangipani
2001 @ LLNL



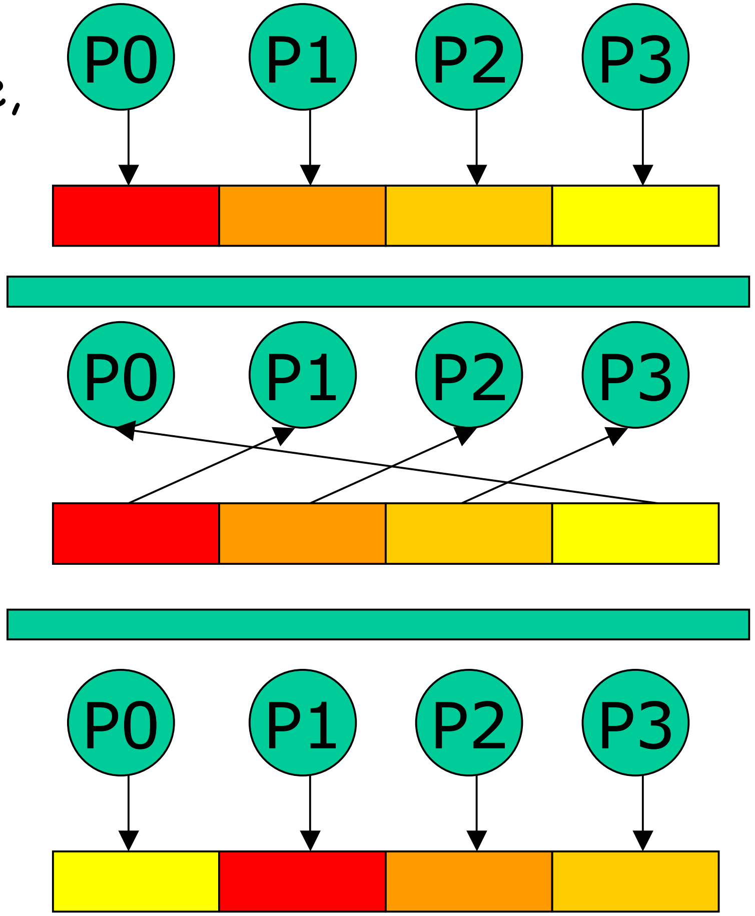
Linux 2.4 Streaming I/O
2002 @ SCS, Zurich



Testing a Distributed File System: Example

Supercomputing Systems

- MPI based distributed test suite, covering
  - meta-data processing
  - data consistency
  - distributed POSIX semantics
- Stress test and correctness test
- 42 individual tests
- also interesting when placing all processes on single node:
  - 2.4.18 ext2 passes 41 (writev's are not atomic)
  - 2.2.18 ext2 passed 26



All processes write into a single file

MPI_Barrier()

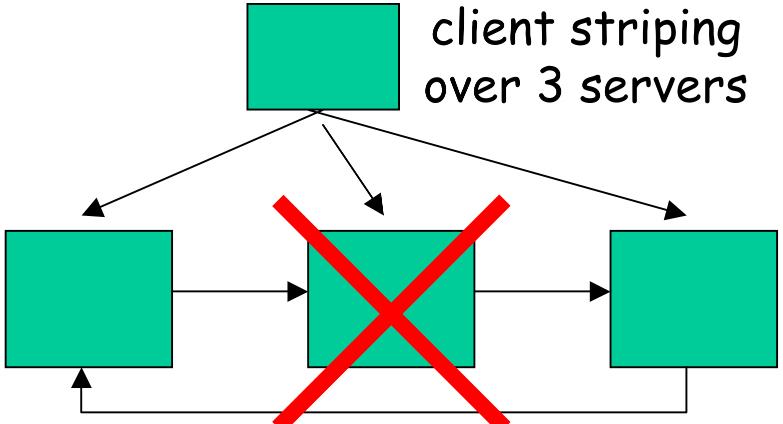
Read data from neighbor on the left

MPI_Barrier()

Write data back to own part of file

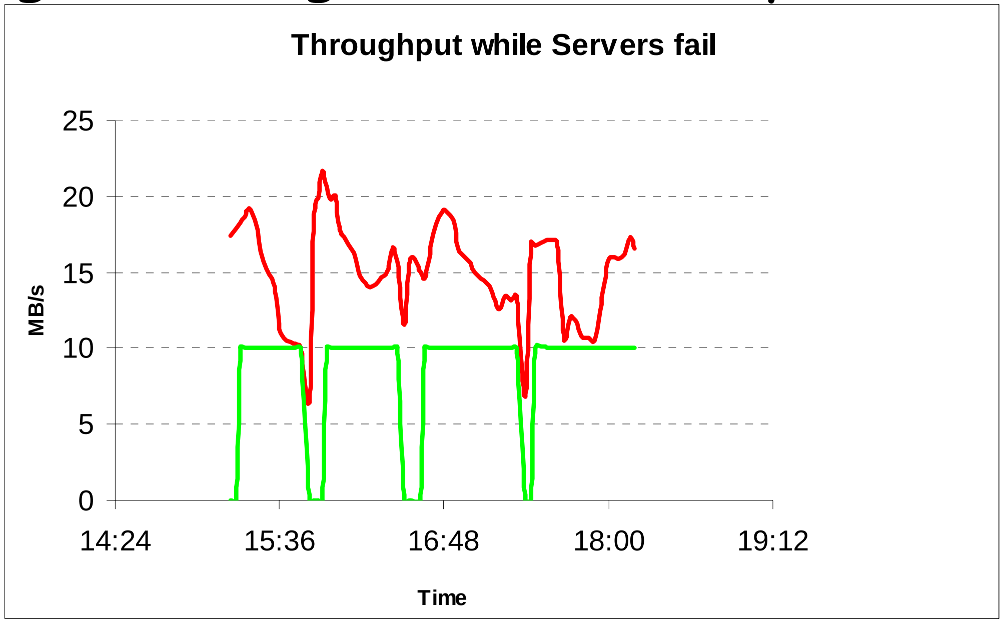
Failure Handling Testing

- Test: Failure of Frangipani Storage Server
- Automated test using
  - power off of server through remote management
  - simulation of Ethernet link failure through switch reconfiguration
- Typical test runs 24h and longer, having a failure every 30 minutes



client striping over 3 servers

3 servers mirroring data to another server



Current Work on File Systems @ SCS

- Apply the tests to the Lustre File System
- Report the bugs, propose fixes https://bugzilla.lustre.org/
- Current (yesterday's) status:
  - » 28 out of 42 work
  - » But: 2 .. 3 fixes per week. Lustre is coming fast!
- Hope to leverage our experience of previous file system work in Lustre